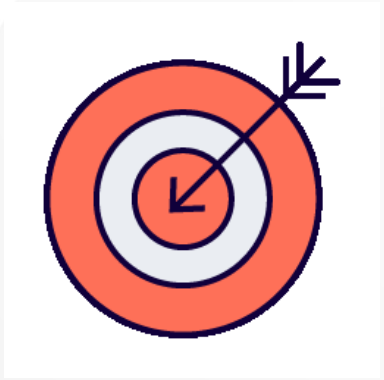
Automation of E & I Processes

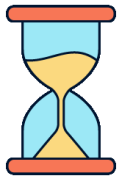
Kerstin Lange
Federal Statistical Office of Germany

Workshop on Statistical Data Editing
31 August – 4 September 2020

Motivation & Aim



- Automation of editing and imputation process
- Increase transparency, more objective
- Standardization
- Faster, save resources and money
- More efficiency



First Steps

- Implement working group with representatives from the Länder
- Aim: development and evaluation of methods
- Investigating existing (machine learning) methods for E & I and assess their potential
- Work out tool requirements as well as advantages and disadvantages

Case Study

New survey: new digital structure of earnings survey starting in 2021

➡ New challenges:

- Larger amount of data: 7 million records from individuals
- More frequent delivery of data: every month
- Current E & I process: manual editing
  - Subject matter expertise
  - Call respondents again

➡ Automation of E & I process is necessary

Test Data

- Similar survey with similar features from 2014
- Features: gross monthly earnings, weekly working hours, paid hours, level of education, demographic features
- Raw material of 500,000 records
- Compare material with new automated imputation methods with manually edited material
- 3% nonresponse
- Main feature: gross monthly earnings

Editing and Imputation with HoloClean

HoloClean:

- Automated tool for error detection and data repairing
- Developed at Stanford University in 2017
- Written in Python, open source

Idea:

- E&I process as supervised machine learning problem
- Deterministic error detection
- Data repairing: model-based imputation; model-building with neural network

Editing and Imputation with HoloClean

Problems:

- Assumption of categorical variables → binning
- Missing functionalities
- Missing documentation
- Runtime

Editing and Imputation with HoloClean

Adjustments:

- Subsample of 10,000 records
- Binning of numerical variables

	Mean gross monthly earnings – manually edited material in Euro	Mean gross monthly earnings – edited with HoloClean in Euro	Difference in %
Overall	2340	2435	4.1
Men	2848	2509	-11.9
Women	1739	2347	35.0

- Runtime: 3 hours (machine with 16GB RAM, 4 kernel CPU)

➡ Not workable for the editing of 7 million records per month

Editing and Imputation with CANCEIS

CANadian Census Edit and Imputation System developed by Statistics Canada

- Donor imputation based on nearest-neighbor imputation methodology
- Already used in structure of earnings survey

Editing and Imputation with CANCEIS

Test data:

- 500,000 records of full time employees

	Mean gross monthly earnings – manually edited material in Euro	Mean gross monthly earnings – edited with CANCEIS in Euro	Difference in %	N
Overall	3226	3223	-0.08	490.795
Men	3477	3473	-0.12	336.551
Women	2677	2678	0.04	154.208
≥ 65 years	3882	3871	-0.29	2.610
Managers	6646	6627	-0.28	45.981

- Runtime: 5 minutes

Editing and Imputation with missForest

missForest:

- R Package
- Imputation of missing values based on a random forest approach
- For continuous and categorical data

Problems:

- No integration of edit rules possible
  ➡ Shoot-out algorithm

Editing and Imputation with missForest

Adjustments:

- Subsample of 10,000 records

	Mean gross monthly earnings – manually edited material in Euro	Mean gross monthly earnings – edited with missForest in Euro	Difference in %
Overall	2502	2482	-0.8
Men	3094	3068	-0.8
Women	1808	1814	0.3

- Satisfying results, a little worse than with CANCEIS
- Still some implausible records

Outlook

- Working on adjusting CANCEIS parameters and improving missForest implementation
- Add more measures for analysis
- Test BANFF as a tool
  - Developed by Statistics Canada
  - Different modules for numeric data